Project: Retail

I. Project Overview:

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Problem Statement:

It is a business critical requirement to understand the value derived from a customer. RFM is a method used for analyzing customer value.

Perform customer segmentation using RFM analysis. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value). Identifying the most valuable RFM segments can capitalize on chance relationships in the data used for this analysis.

Data set Description:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

II. Analysis:

Data Exploration:

1. Performed a preliminary data inspection and Data cleaning

    a. Check for missing data and formulate apt strategy to treat them.

In the Dataset 25% of missing data in Customer ID variable and 0.26% of data missing in Description Variable. Hence decided to drop the Customer ID missing data.

b. Are there any duplicate data records? Remove them if present.
5225 observations were duplicated and dropped all the duplicates

c. Perform Descriptive analytics on the given data.
Performed the Descriptive analysis as given below

:

	Quantity	UnitPrice	CustomerID
count	401604.000000	401604.000000	401604.000000
mean	12.183273	3.474064	15281.160818
std	250.283037	69.764035	1714.006089
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13939.000000
50%	5.000000	1.950000	15145.000000
75%	12.000000	3.750000	16784.000000
max	80995.000000	38970.000000	18287.000000

2. Cohort Analysis: A cohort is a group of subjects who share a defining characteristic. We can observe how a cohort behaves across time and compare it to other cohorts.
   a. Create month cohorts and analyse active customers for each cohort.
   To Create month Cohorts, transformed the date column into month and year and calculated customer wise first transaction month as Cohort Month. And to analyse the active customers for each cohort calculated the cohort index as the difference between the cohort month and Transaction month . Then calculated the cohort active users by grouping cohort month and index and calculating the number of unique customers in the group

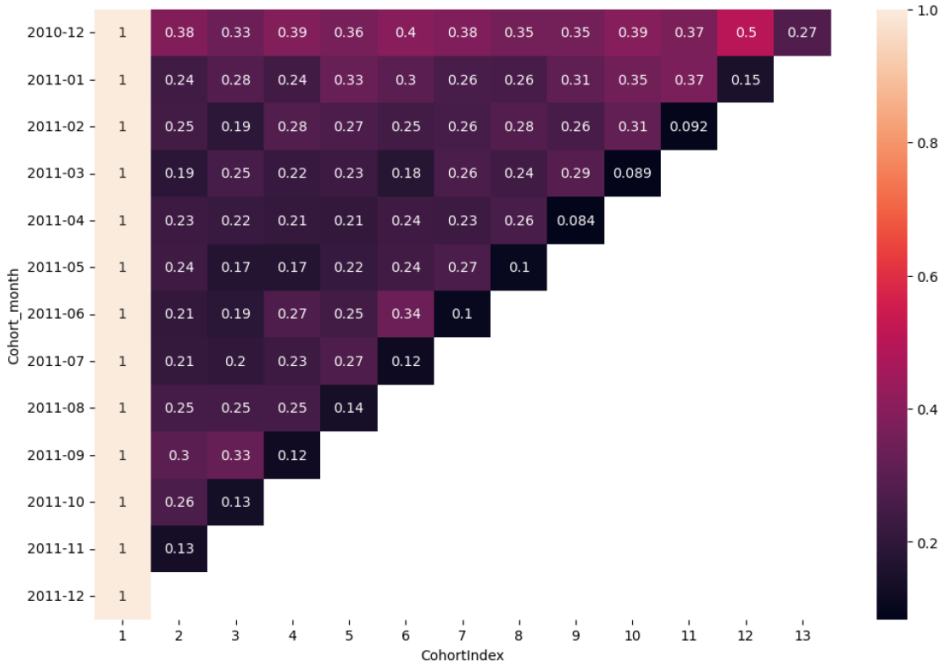
CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
Cohort_month													
2010-12-01	948.0	362.0	317.0	367.0	341.0	376.0	360.0	336.0	336.0	374.0	354.0	474.0	260.0
2011-01-01	421.0	101.0	119.0	102.0	138.0	126.0	110.0	108.0	131.0	146.0	155.0	63.0	NaN
2011-02-01	380.0	94.0	73.0	106.0	102.0	94.0	97.0	107.0	98.0	119.0	35.0	NaN	NaN
2011-03-01	440.0	84.0	112.0	96.0	102.0	78.0	116.0	105.0	127.0	39.0	NaN	NaN	NaN
2011-04-01	299.0	68.0	66.0	63.0	62.0	71.0	69.0	78.0	25.0	NaN	NaN	NaN	NaN
2011-05-01	279.0	66.0	48.0	48.0	60.0	68.0	74.0	29.0	NaN	NaN	NaN	NaN	NaN
2011-06-01	235.0	49.0	44.0	64.0	58.0	79.0	24.0	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	191.0	40.0	39.0	44.0	52.0	22.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	167.0	42.0	42.0	42.0	23.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	298.0	89.0	97.0	36.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	352.0	93.0	46.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

b. Also Analyse the retention rate of customers. Comment.
Retention rate is calculated by dividing the active customers by cohort size(no. of customers in each customer)

:

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
Cohort_month													
2010-12	100.0	38.2	33.4	38.7	36.0	39.7	38.0	35.4	35.4	39.5	37.3	50.0	27.4
2011-01	100.0	24.0	28.3	24.2	32.8	29.9	26.1	25.7	31.1	34.7	36.8	15.0	NaN
2011-02	100.0	24.7	19.2	27.9	26.8	24.7	25.5	28.2	25.8	31.3	9.2	NaN	NaN
2011-03	100.0	19.1	25.5	21.8	23.2	17.7	26.4	23.9	28.9	8.9	NaN	NaN	NaN
2011-04	100.0	22.7	22.1	21.1	20.7	23.7	23.1	26.1	8.4	NaN	NaN	NaN	NaN
2011-05	100.0	23.7	17.2	17.2	21.5	24.4	26.5	10.4	NaN	NaN	NaN	NaN	NaN
2011-06	100.0	20.9	18.7	27.2	24.7	33.6	10.2	NaN	NaN	NaN	NaN	NaN	NaN
2011-07	100.0	20.9	20.4	23.0	27.2	11.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08	100.0	25.1	25.1	25.1	13.8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09	100.0	29.9	32.6	12.1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10	100.0	26.4	13.1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11	100.0	13.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Cohort Retention rate heat map



Here, We have 13 cohorts for each month and 13 cohort indexes. If we see in 2010-12 cohort Month in 12th Cohort Index, we see the red shade with 0.5 value which means that 50% of cohorts that signed in December 2010 were active 12 months later.

3. Build a RFM model – Recency Frequency and Monetary based on their behaviour.
   a. Calculate RFM metrics.
      i. Recency as the time in no. of days since last transaction

ii. Frequency as count of purchases done
iii. Monetary value as total amount spend

Built an RFM model by calculating the Customer wise Recency, Frequency and Monetary

	CustomerID	Recency	Frequency	Monetry
0	12346.0	325	2	0.00
1	12347.0	1	7	4310.00
2	12348.0	74	4	1797.24
3	12349.0	18	1	1757.55
4	12350.0	309	1	334.40

b. Build RFM Segments.

I. Calculated the scores for Recency, Frequency and Monetary by diving them into quratiles

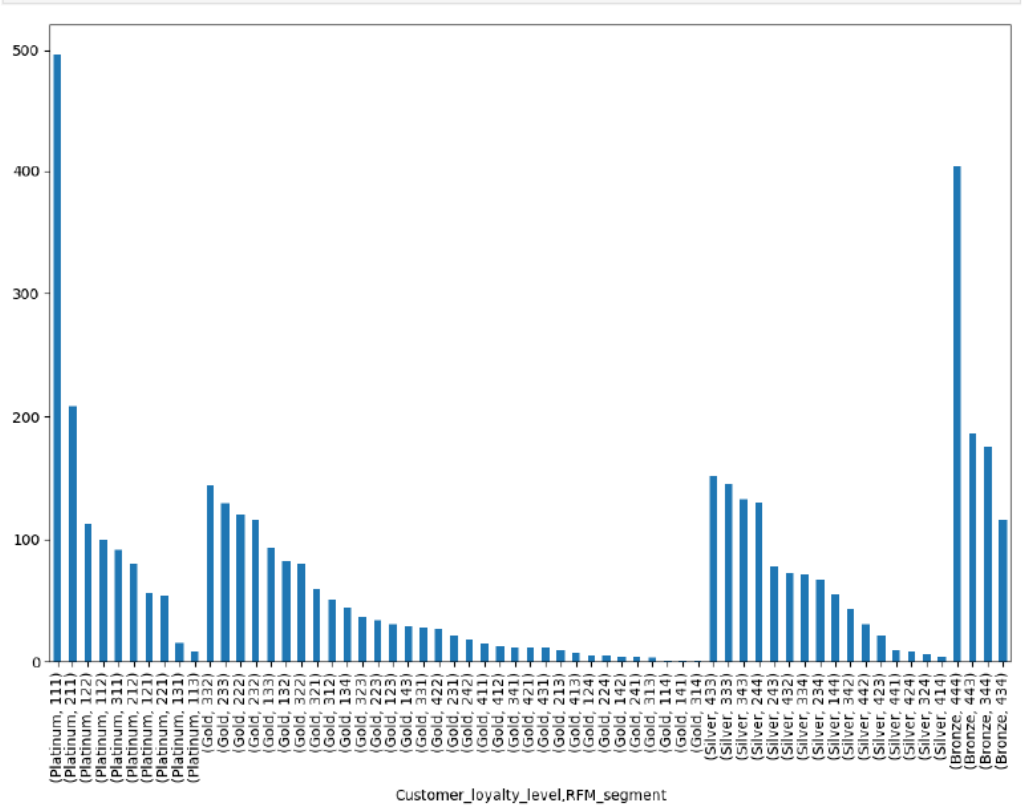
	CustomerID	Recency	Frequency	Monetry	R	F	M
0	12346.0	325	2	0.00	4	3	4
1	12347.0	1	7	4310.00	1	1	1
2	12348.0	74	4	1797.24	3	2	1
3	12349.0	18	1	1757.55	2	4	1
4	12350.0	309	1	334.40	4	4	3

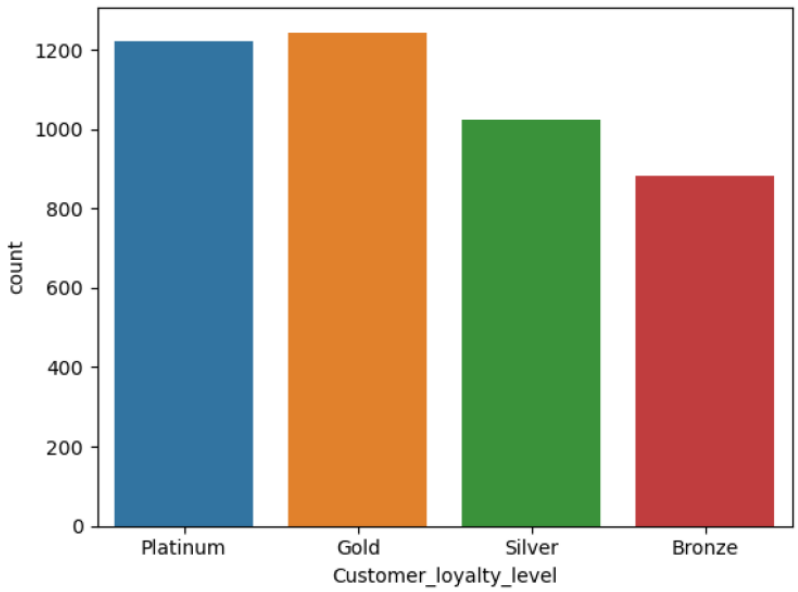
II. Combined the ratings as strings to calculate RFM segment and added them together to calculate the RFM score

	CustomerID	Recency	Frequency	Monetry	R	F	M	RFM_segment	RFM_Score
0	12346.0	325	2	0.00	4	3	4	434	11
1	12347.0	1	7	4310.00	1	1	1	111	3
2	12348.0	74	4	1797.24	3	2	1	321	6
3	12349.0	18	1	1757.55	2	4	1	241	7
4	12350.0	309	1	334.40	4	4	3	443	11
5	12352.0	35	11	1545.41	2	1	2	212	5
6	12353.0	203	1	89.00	4	4	4	444	12
7	12354.0	231	1	1079.40	4	4	2	442	10
8	12355.0	213	1	459.40	4	4	3	443	11
9	12356.0	22	3	2811.43	2	3	1	231	6

c. Analyse the RFM Segments by summarizing them and comment on the findings

Assigned the Loyalty level to each customer based on RFM scores as Platinum, Gold, Silver and Bronze and there distribution based on each RFM segment is as given below





The above Graph shows the Customer Grouping based on their Loyalty level based on recency, frequency and Monetry

28% of Customers are having Platinum Level who are very good in terms of recency, frequency and Monetry

28% of Customers are having Gold level who are fair in terms of recency, frequency and Monetry

23.4% of Customers are in Silver level who are not very recent, frequent and monetry

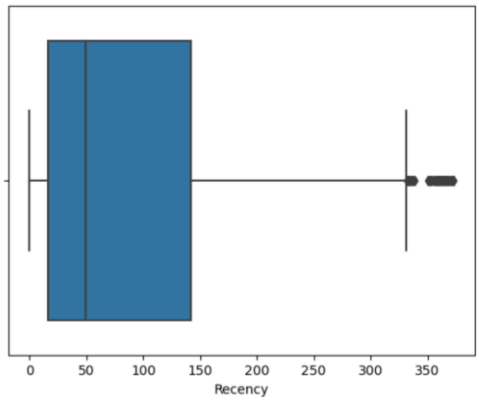
20% of Customers are Bronze level who are not recent, frequent and monetary

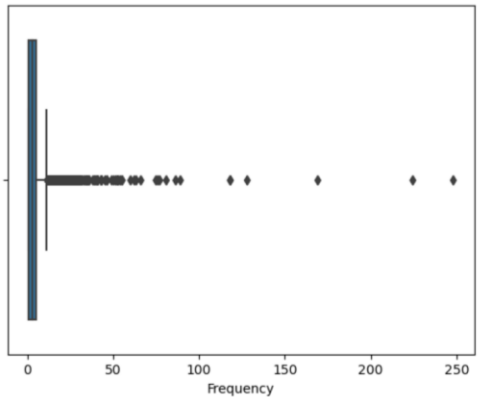
III. Algorithms and Techniques:

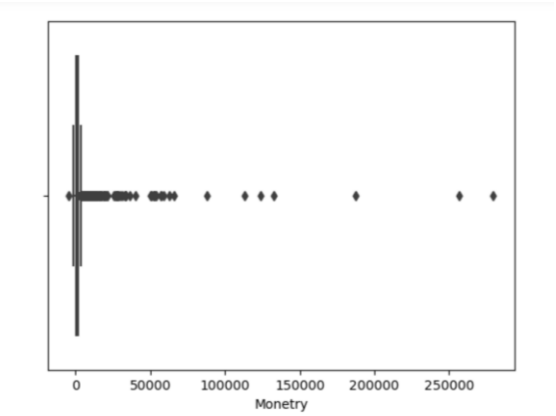
This is a clustering problem, hence building the model using K means clustering. To calculate the optimum number of cluster will plot the elbow curve and derive the optimum number of clusters

Data Preprocessing:

Checked for the outliers using the boxplot for Recency, Frequency and Monetary variables



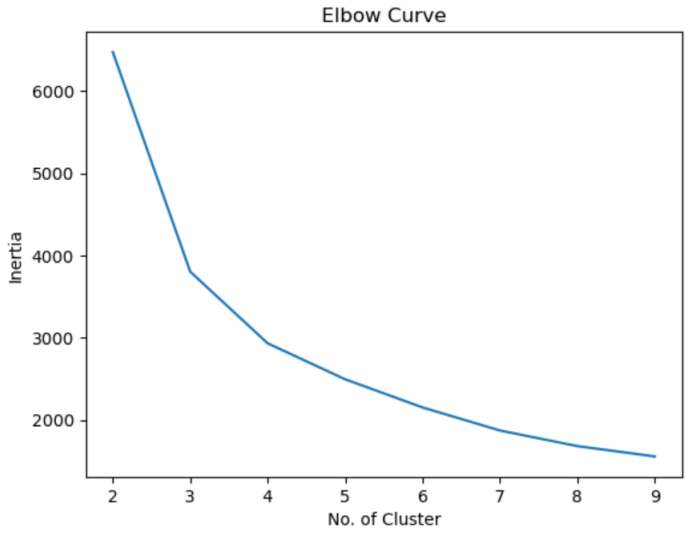




and treated the outliers from Monetary, Frequency and Recency columns by dropping it and then data is scaled

Model Building:

Built a K means Clustering model and checked the optimum number of cluster using elbow curve method by plotting the no. cluster vs. inertia



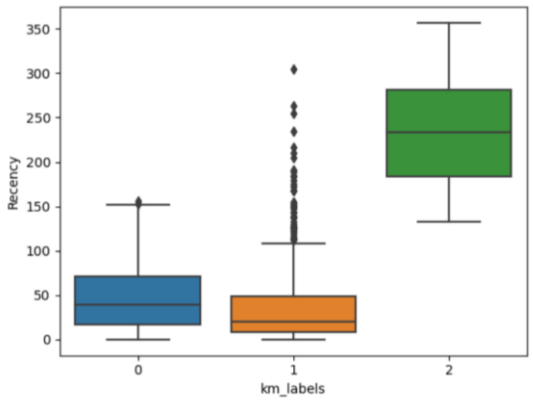
IV. Results:

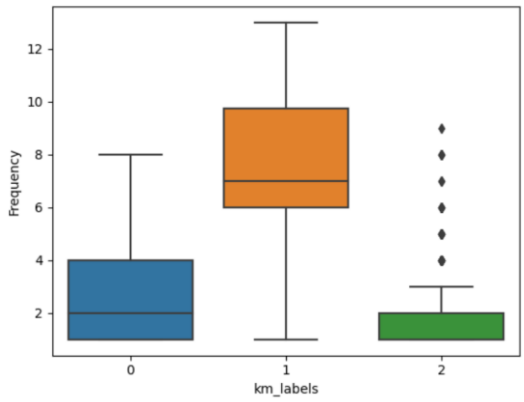
With the resultant elbow curve, identified that optimum number of cluster is 3. Hence built the K means cluster model with 3 no. of clusters.

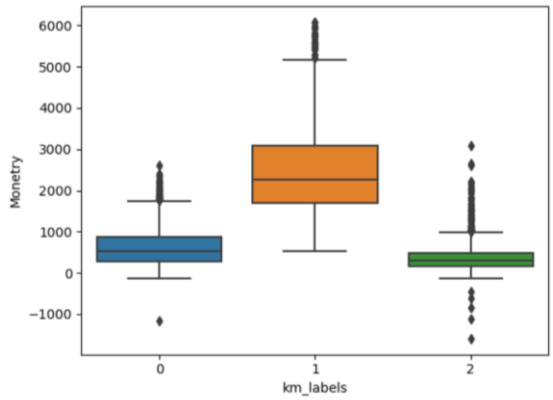
	CustomerID	Recency	Frequency	Monetry	km_labels
0	12346.0	325	2	0.00	2
1	12347.0	1	7	4310.00	1
2	12348.0	74	4	1797.24	0
3	12349.0	18	1	1757.55	0
4	12350.0	309	1	334.40	2
...	...	...	...	...	...
4366	18278.0	73	1	173.90	0
4367	18280.0	277	1	180.60	2
4368	18281.0	180	1	80.82	2
4369	18282.0	7	3	176.60	0
4371	18287.0	42	3	1837.28	0

3910 rows × 5 columns

Plotted the Box plot for all 3 clusters with respect to Recency, Frequency and Monetary as given below







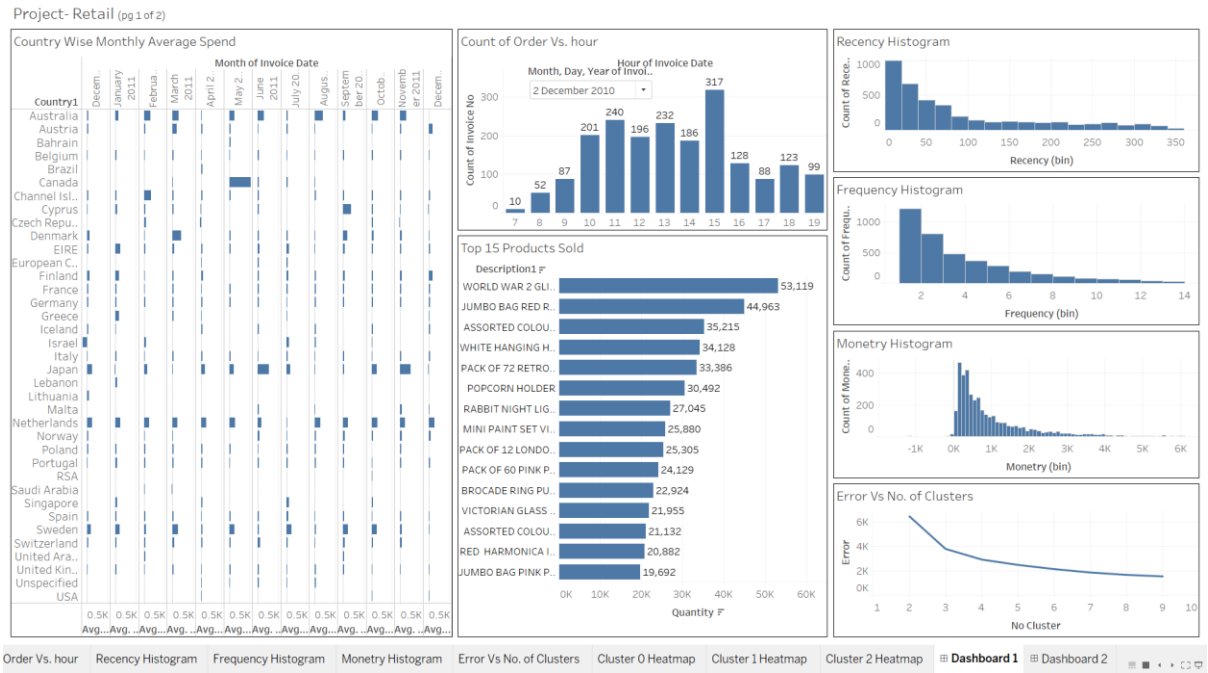
V. Conclusion:

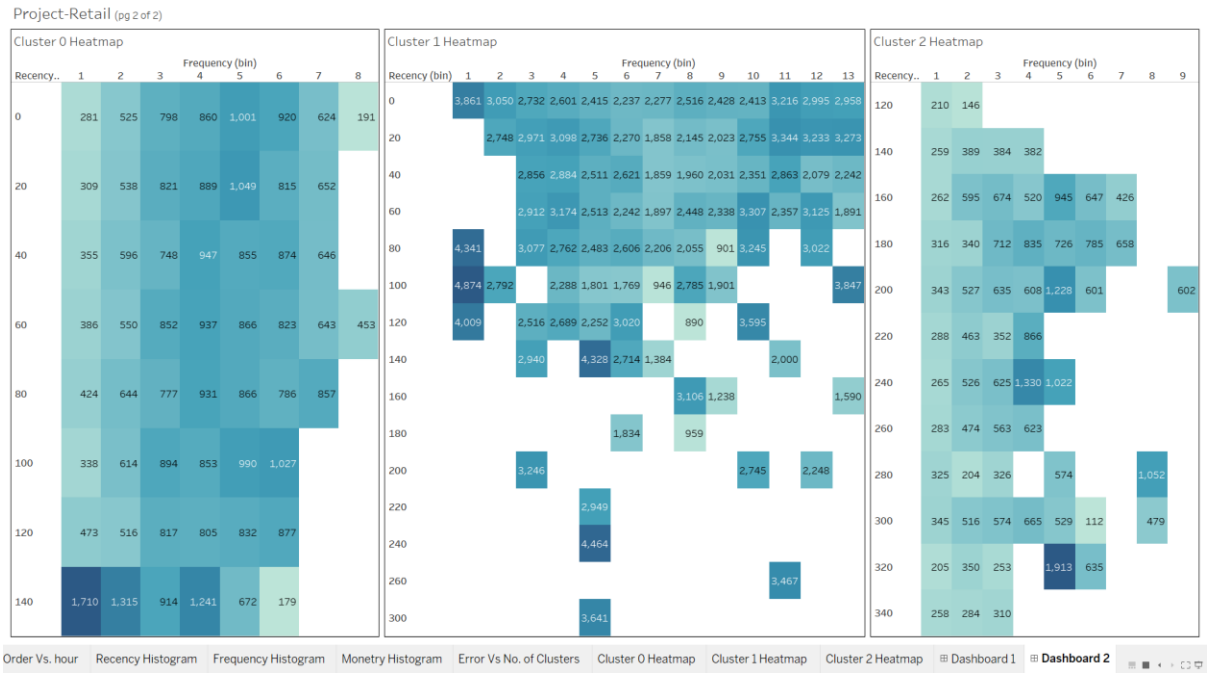
Cluster 0 and Cluster 1 customers are Recent, Frequent and spends well when compared with Cluster 2 customers.

VI. Tableau Dashboard:

The Dashboard has following visuals

a) Country-wise analysis to demonstrate Average spend. Using a bar chart to show monthly figures.
b) Bar graph of top 15 products which are mostly ordered by the users to show the number of products sold.
c) Bar graph to show the count of orders Vs. hours throughout the day. What are the peak hours per your chart?
d) Plot the distribution of RFM values using histogram and frequency-charts.
e) Plot error(cost) vs no of clusters selected
f) Visualize to compare the RFM values of the clusters using heatmap




Project-Retail (pg 2 of 2)
Cluster 0 Heatmap
Recency..
Frequency (bin)
0 281 525 798 860 1,001 920 624 191
20 309 538 821 889 1,049 815 652
40 355 596 748 947 855 874 646
60 386 550 852 937 866 823 643 453
80 424 644 777 931 866 786 857
100 338 614 894 853 990 1,027
120 473 516 817 805 832 877
140 1,710 1,315 914 1,241 672 179
Cluster 1 Heatmap
Recency (bin)
Frequency (bin)
0 3,861 3,050 2,732 2,601 2,415 2,237 2,277 2,516 2,428 2,413 3,216 2,995 2,958
20 2,748 2,971 3,098 2,736 2,270 1,858 2,145 2,023 2,755 3,344 3,233 3,273
40 2,856 2,884 2,511 2,621 1,859 1,960 2,031 2,351 2,863 2,079 2,242
60 2,912 3,174 2,513 2,242 1,897 2,448 2,338 3,307 2,357 3,125 1,891
80 4,341 3,077 2,762 2,483 2,606 2,206 2,055 901 3,245 3,022
100 4,874 2,792 2,288 1,801 1,769 946 2,785 1,901 3,847
120 4,009 2,516 2,689 2,252 3,020 890 3,595
140 2,940 4,328 2,714 1,384 2,000
160 3,106 1,238 1,590
180 1,834 959
200 3,246 2,745 2,248
220 2,949
240 4,464
260 3,467
300 3,641
Cluster 2 Heatmap
Recency..
Frequency (bin)
120 210 146
140 259 389 384 382
160 262 595 674 520 945 647 426
180 316 340 712 835 726 785 658
200 343 527 635 608 1,228 601 602
220 288 463 352 866
240 265 526 625 1,330 1,022
260 283 474 563 623
280 325 204 326 574 1,052
300 345 516 574 665 529 112 479
320 205 350 253 1,913 635
340 258 284 310
Order Vs. hour
Recency Histogram
Frequency Histogram
Monetry Histogram
Error Vs No. of Clusters
Cluster 0 Heatmap
Cluster 1 Heatmap
Cluster 2 Heatmap
Dashboard 1
Dashboard 2